Affect Effects in Social Media: Relation Between Automated Emotion Detection, User-Reported Emotional Responses, and Post Interaction Behavior

Aryn Pyke[1,2], Darielicia Finley[2], Aaron Bonner[2], Iain Cruickshank[1]

[1] Army Cyber Institute, West Point NY, USA
[2] West Point (United States Military Academy), West Point NY, USA
aryn.pyke@westpoint.edu

Abstract. We investigated the relationship between users' emotional reactions - both nature and intensity - and their interaction behavior with social media posts (like/dislike, share, comment). On a simulated social media interface, participants had the opportunity to interact with text posts designed to provoke positive, negative or emotionally neutral responses. Interaction level was highest for positive posts, then negative posts, and lowest for neutral posts. Subjects' baseline emotional states were skewed toward positive emotions which may have impeded some negative emotions from reaching threshold for action on negative posts. Posts were more likely to be liked (81% of positive posts) or disliked (63% of negative posts) than shared (24% of polar posts) or commented on (17% of polar posts). Users also self-reported the intensity of their own emotional responses to posts on each of four scales (roughly: joy, pleased, sad, and mad). Overall level of interaction and sharing likelihood were predicted by their joy, sad and mad ratings. All four scales predicted like/dislike likelihood. However, only the joy (and marginally mad) ratings predicted commenting behavior. More active forms of emotion may be needed to motivate the extra effort to comment. We also compared users' reported emotional responses with emotional responses predicted from post texts via automatic emotion detection (Chat GPT-3.5) and found them to be well correlated (rs > .56) in spite of the posts being designed for students at a particular institution. Although not suited to capture individual differences, automatically inferred emotional ratings seem a good proxy for aggregate measures.

Keywords: social media interactions, emotion, affect, emotion detection, Chat GPT.

1 Introduction

When a user reads a post on social media, there are several factors that may affect the likelihood the user engages with the post (e.g., likes/dislikes, comments, shares), and thereby contributes to its further propagation/dissemination. For example, users may be more likely to share a post if the content in the post reflects their existing beliefs ***(confirmation bias***; Kim & Dennis, 2018). Additionally, users are more likely to like/share

posts from a source they trust (Buchanan & Benson, 2019; Kim & Dennis, 2018) or to whom they have strong social ties (Apuke & Oman, 2020).

In the current research, we investigated the relationship between users' emotional reaction (both nature and intensity) and their engagement with social media posts. We were particularly interested in the impact of the emotional state induced by the post, but a user's initial emotional state prior to reading content can also have an impact. A prior study reported that being in a more intense emotional state, whether negative or positive, can impair one's ability to recognize the inaccuracy of false claims (i.e., 'fake news'; Martel, Pennycook & Rand, 2020). In the current research, we had participants report their baseline emotional state at the start of the study, as well as their emotional state after each post, so that we could determine the relation between emotional state and post interaction behavior.

In research involving posts scraped from social media, there are typically no data on the individual emotional reactions of the users. Instead, natural language processing algorithms can be applied to the text of a post to categorize or quantify the affective nature of a post. For example, **sentiment analysis** can categorize the emotional tone of a post as positive, negative or neutral. More sophisticated algorithms (e.g., ***emotion detection***) can quantify the degree a post reflects different emotions (like happiness, sadness, and anger). On Twitter, messages with content that these algorithms suggest is emotionally charged tended to be retweeted more often and more quickly than posts that the algorithms rated as more neutral. Another aim of this research was to compare the emotional content inferred by automated emotion detection against the emotional reactions to the posts self-reported by the participants themselves.

2 Current Research

On a simulated social media interface (Fig. 1), participants had the opportunity to interact with posts that were designed to provoke either a positive, negative or emotionally neutral response. We expected that users would interact more with positive and negative posts than neutral posts, but it was an open question if there would be a difference between positive and negative posts. Further, to get a more nuanced understanding of how the nature and intensity of users' emotional responses can impact interaction behavior, regardless of the intended response to a post, we also had users self-report the intensity of their own individual emotional response to a post in each of four scales (Fig. 2): i) Joy/Excitement/Elation; ii) Pleased/Contented; iii) Sad/Disappointed/Worried; iv) Mad/Disgusted/Shocked & Appalled. As a shorthand, we will often refer to each scale just by its first descriptor (e.g., joy, pleased, sad, and mad).

The inclusion of two positive (i, ii) and two negative scales (iii, iv) was motivated to try to distinguish more passive emotional responses (positive: pleased, negative: sad) from more active ones (positive: excited, negative: mad). To support this view, within the negative scale, evidence suggests that anger produces an approach motivation whereas some other negative emotional responses like anxiety can be associated with avoidance behavior (Carver & Harmon-Jones, 2009). This distinction was important because we hypothesized that active/approach forms of emotions are more likely to

motivate action, and thus that the intensity of the more active scale would be better predictors of interaction behavior.

In addition to investigating the relation between emotional response and interaction behavior, another aim of this research was to compare human emotional responses to the posts with the emotional characterization of the posts that would be inferred by an automated emotion detection process (using Chat GPT-3.5).

3 Human Study Methods

3.1 Participants

Participants (N=91, 73% male, Mean Age = 20.2 years, SD = 1.3) were college students from the psychology pool at West Point (United States Military Academy) and received course credit for their participation.

3.2 Materials and Procedure

The study was conducted on-line (implemented in Qualtrics™) and participants had the opportunity to interact with hypothetical posts in a simulated social media interface that afforded the ability to like/dislike, comment on, and share posts (see Fig. 1). Participants were told to pretend that they were surfing on a social media site for only current students at their institution, and that it was used by students to share information about possible upcoming changes at the college which had been heard through the grapevine. They were asked to do their best to respond as they would if they saw these posts on their own platform (but also told that their responses would be anonymous to the researchers and could not be linked back to them).

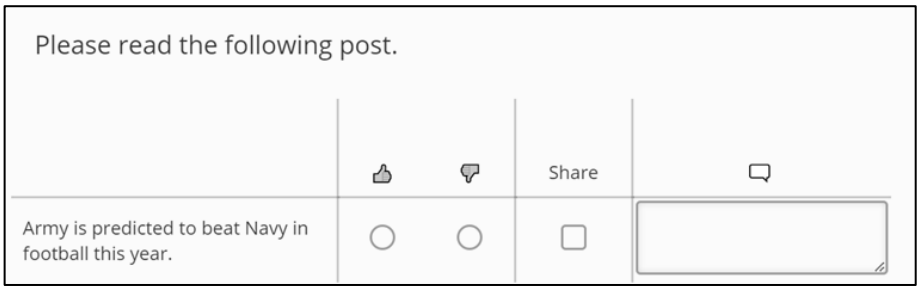

Fig. 1. An example post illustrating the simulated social media interface for this study which allowed subjects to like/dislike, comment on, and share posts.

Each post was just plain text (no pictures, links nor hashtags), and to control for source effects, no information about the poster was provided (beyond the implication that it was hypothetically written by another student at their institution). Posts were designed with the intention to provoke either positive, neutral or negative emotional reactions in this population of students. An example of a positive post was: “Winter

break may be extended by 2 days". The corresponding negative post was: "Winter break may be shortened by 2 days". However, a given subject would see only see either the negative or positive version of each positive-negative post pair (counter-balanced across participants). Neutral posts tended to describe the current state-of-affairs rather than to forecast any potential positive or negative change, for example "The mail room will have the same hours as last semester".

There were 4 different sets of posts (A1, A2, B1, B2) each with 5 positive and 5 negative posts (from different pairs) and 10 neutral posts. Each participant was assigned randomly to see one set of posts. Thus, participants read and had the opportunity to interact with 20 posts each: 5 positive, 5 negative and 10 neutral. Every other post, starting with the first post was a neutral post to provide a buffer (emotional palette cleanser) between the intended emotion-inducing posts (positive and negative). A single random order of the positive and negative posts was used for each set with the constraint that no more than two posts of a given polarity could be "in a row" (with neutral posts interspersed).

To validate the intended manipulation of post type (positive, neutral, negative), and to collect information on individual emotional responses, after engaging with each post, participants rated their emotional reactions on a 4-scaleal Likert scale: i) Joy/Excitement/Elation; ii) Pleased/Contented; iii) Sad/Disappointed/Worried; iv) Mad/Disgusted/Shocked & Appalled (see Fig. 2). At the outset of the study, participants also rated their initial baseline emotional state using the same scales.

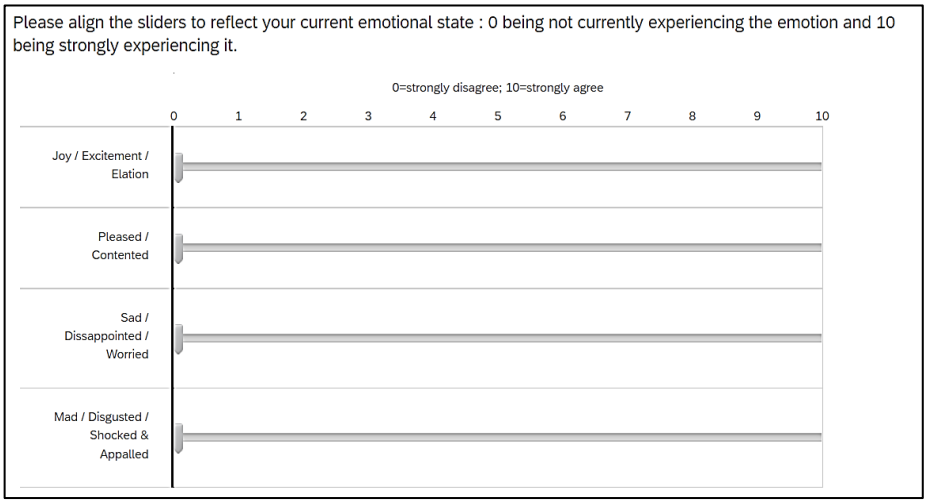

Fig. 2. Likert scales allowing subjects to report their emotional responses to a post.

4 Automated Emotion Detection Method

One aim of this research was to compare human emotional responses to the posts with the emotional characterization of the posts that would be provided by an automated

emotion detection process. Chat GPT-3.5 was used to infer emotional content from the post texts and to produce output using the same emotion categories/scales used by human subjects (Fig. 2). Here is an example of the prompt:

Prompt: "'You are a Cadet at the United States Military Academy, at West Point. The following is a social media post relevant to being a cadet at the Academy. Please give a rating from 0 to 10 for how much the post inspires an emotion within you, with 0 being not at all and 10 being strongly. The emotions you need to evaluate are "Joy/Excitement/Elation", "Pleased/Contented", "Sad/Disappointed/Worried", and "Mad/Disgusted/Shocked and Appaled".

For each emotion category please give a score from 0 to 10 and output the scores in the form {emotion:score}.

post: Winter break may be shortened by 2 days.

emotions:"'

GPT-3.5: "' {Joy/Excitement/Elation: 8, Pleased/Contented: 7, Sad/Disappointed/Worried: 2, Mad/Disgusted/Shocked and Appalled: 1}"'

The emotional characterization produced by the method above was compared with the self-reported emotional characterization of the posts by human participants.

5 Results

5.1 Human and Chat GPT Emotional Ratings

As a manipulation check, Fig. 3(left panel) illustrates the mean emotional ratings by subjects for each of the post types. As expected, negative posts elicited higher sad and mad ratings than positive posts, and positive posts elicited higher joy and pleased ratings than negative posts. Fig. 3(right panel) illustrates the ratings inferred by Chat GTP, averaged across the 20 posts within each post type. Although the ratings inferred by Chat GPT tended to be higher, the patterns are remarkably similar to the human ratings in the aggregate.

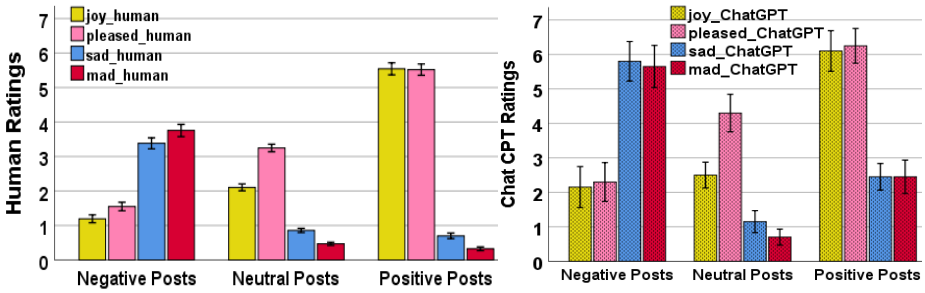

Fig. 3. Mean ratings in each of our four emotional scales by post type from human self-reports (left panel) and inferred from post texts by Chat GPT-3.5 (right panel). Rating scales were from 0 (lowest intensity) to 10 (highest intensity). Error bars are standard error.

Across all 60 post items (20 positive, 20 negative, 20 neutral) we did a correlation analysis for each of our emotion scales to find the strength of the relation between the human ratings (aggregated by post) and Chat GPT emotion ratings. Correlations were all significant ($p<.001$) and fairly high: $r=.600$ for joy, $r=.562$ for pleased, $r=.663$ for sad; $r=.617$ for mad.

Within the ratings across the 60 post items there were also significant correlations across scales (all $p<.001$). For example, within the human ratings, joy was correlated with pleased ($r=.915$); and negatively correlated with sad ($r=-.610$), and mad ($r=.583$); pleased was negatively correlated with sad ($r=.--.750$) and mad ($r=-.746$), and mad was correlated with sad ($r=.938$).

5.2 Effect of Post Type on Interaction Behavior

For the different post types, Fig. 4 illustrates the proportion of posts that were liked, disliked, shared and commented upon. Interestingly, there were a few likes of posts intended to be negative and a few dislikes of posts intended to be positive. Liking and disliking interactions were clearly more common than shares and comments.

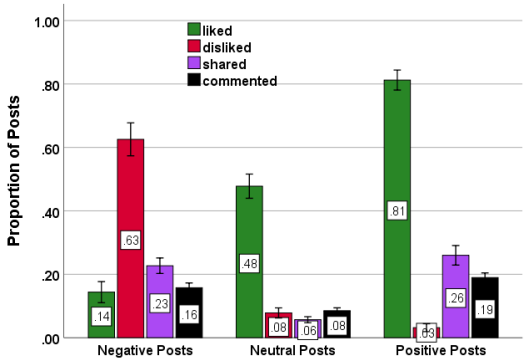

Fig. 4. Proportion of interactions of various types by post type. Error bars are standard error.

Note that the interactions in Fig. 4 are not mutually exclusive - for a given post a user may like, and share and comment. To capture such 'compound' engagements, for each user's interaction with a post we created an interaction score which was the sum of 1 point for a like/dislike, 1 point for a comment, and 1 point for sharing, for a max of 3 points. As depicted in Fig. 5, a repeated measures ANOVA revealed a main effect of post type on interaction scores, $F(2,180) = 93.72$, partial-eta$^2 = .51$, $p<.001$. Pairwise tests with a Bonferroni correction confirmed that scores for positive posts (1.29) were greater than those for negative posts (1.15, $p = .001$), which in turn were greater than those for neutral posts (.70, $p<.001$).

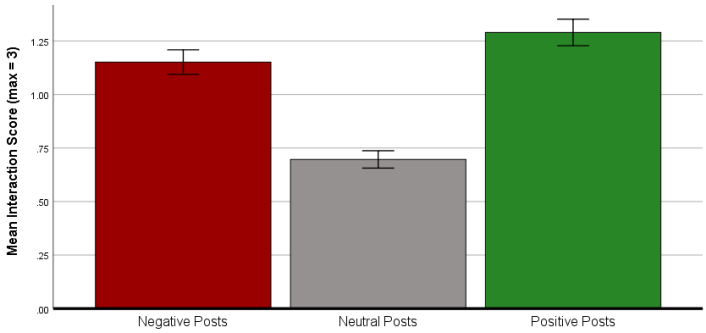

Fig. 5. Mean interaction scores by post type. Interaction score is the sum of 1 point for a like/dislike, 1 post for a comment, and 1 point for sharing, for a max of 3 points. Error bar is STE.

5.3 Effect of Human Baseline Emotional State on Interaction Behavior

In order of intensity, the mean baseline self-reported emotional scores out of 10 at the outset of the study were: pleased (M=5.6, SD=2.4); joy (M=4.6, SD=2.7); sad (M=2.9, D=2.6); and mad (M=1.0, SD=1.6). To determine if baseline emotional state was predictive of interaction behavior, we correlated each emotional scale with the interaction scores for positive, negative and neutral posts. Only the baseline intensity of the Sad/Disappointed/Worried scale was predictive of interactions. It predicted increased interactions with positive posts (r = .334, p=.001) and negative posts (r =.315, p=.002), but not neutral posts (r=.022, p=.835).

5.4 Effect of Human Emotional Response to Posts on Interaction Behavior

We expected that posts which elicited higher emotion ratings would promote greater interaction behavior. To test this prediction, we compared the human emotional ratings for posts with different interaction levels. Each interaction type (like/dislike, comment, share) gets 1 point, so no interaction is level 0, all three interactions for a single post a max of 3 points. Fig. 6 illustrates that emotional responses were relatively low on posts which elicited no interactions, but that increasing emotional intensity was associated with increased interaction levels. Intensity of positive emotions tends to promote increased interaction for positive and neutral posts while intensive of negative emotions tends to promote increased interaction for negative posts.

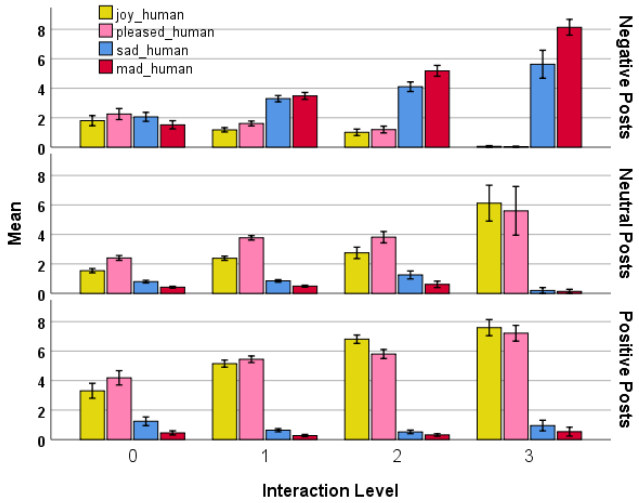

Fig. 6. Mean human emotional ratings in each of our four emotional scales by interaction level and post type. Rating scales were from 0 (lowest intensity) to 10 (highest intensity). Error bars are standard error.

In terms of correlations, interaction level is predicted by joy (r=.431, p=.001), sad (r=.323, p=.012), and mad (r=.357, p=.005). Comment proportion is only predicted by joy (r=.330, p=.010; though mad is marginal, r=.223, p=.087). Share proportion is predicted by joy (r=.293, p=.023), sad (r=.317, p=.014), and mad (r=.335, p=.009). Like proportion is predicted by all four scales (p<.001): joy (r=.865), pleased (r=.933), sad (r=-.730), and mad (r=-.753). And the same is true for dislike proportion (p<.001): joy (r=-.629), pleased (r=-.782), sad (r=.906), and mad (r=.956).

6 Discussion and Conclusions

This research investigated the relation between emotional responses behavioral interactions with posts (like/dislike, comment, share). Posts were more likely to be liked (81% of positive posts) or disliked (63% of negative posts) than shared (24% of polar posts) or commented on (17% of polar posts). Overall, users' level of interaction was highest for positive posts, followed by negative and then neutral posts.

Users had a more positive reaction to neutral posts than we had expected, but this may reflect satisfaction with the status quo because these posts often described current campus practices (e.g., hours that services are available). We had also expected users to be most reactive to negative posts, however this could be an emotional threshold issue. Users' intensity of positive emotional responses (joy and pleased) to positive posts was higher than their intensity of negative emotional responses (sad and mad) to negative posts. This was unexpected as the positive and negative posts were intended to be 'matched'- i.e., having winter break lengthened by two days (positive) versus having winter break shortened by two days (negative). A possible contributing factor

to the relatively lower intensity of negative reactions to negative posts is that users' mean baseline emotional state at the start of the study was skewed toward positive emotions.

Consistent with Martel, Pennycook and Rand (2020), this baseline emotional state at the start of the study in and of itself can predict interaction behavior. However, in the current study, this impact - increased engagement with positive and negative but not neutral posts - was only related to the intensity of a negative emotion (Sad/Disappointed/Worried) and not also a positive one. We expected that joyful and mad states might be more motivating and predictive. However, in a sad state one may be motivated to seek social connection by engaging with social media.

In terms of the relation between users' emotional reactions to the specific posts and their types and levels of interaction, commenting behavior was only predicted by the emotion scale of joy (and marginally mad), which is consistent with our hypothesis that more active forms of positive and negative emotion (vs. pleased and sad, respectively) might motivate more engagement. In contrast to liking, disliking and sharing, which require a single button press, composing a comment requires more effort, which may explain why it is predicted only by more motivating forms of emotional reactions to a post (joy and marginally anger). In this same vein, the 'passive' pleased scale failed to significantly predict shares or overall interaction level. However, all four scales predicted the likelihood of likes and dislikes.

In comparing human self-reported emotional reactions to the posts with automated ratings based on the post texts via Chat GPT, we discover that the response patterns were similar in the aggregate and the correlations were relatively high for all four scales (all above r=.56). This was somewhat surprising as the posts had been designed for students at a particular institution. This correspondence between human and automatically inferred ratings bodes well for the use of automated ratings to predict interactions (at least in the aggregate, as they could not capture individual differences in responses within our participants).

Acknowledgments. This research was supported in part by the Office of Naval Research under the MURI: Persuasion, Identity, & Morality in Social-Cyber Environments Research Grant #N000142112749. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the US Army, the Department of Defense or the US Government.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

Apuke, O. D., & Omar, B. (2020). Modelling the antecedent factors that affect online fake news sharing on COVID-19: the moderating role of fake news knowledge. Health Education Research, 35(5), 490-503.

Buchanan, T., & Benson, V. (2019). Spreading Disinformation on Facebook: Do Trust in Message Source, Risk Propensity, or Personality Affect the Organic Reach of "Fake News"? Social Media + Society.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. Psychological bulletin, 135(2), 183.

Kim, Antino and Dennis, Alan R., Says Who? The Effects of Presentation Format and Source Rating on Fake News in Social Media (August 16, 2018). MIS Quarterly, Vol. 43, No. 3, pp. 1025–1039 (2019).

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. Cognitive research: principles and implications, 5, 1-20.